# Sequence-Based Association and Selection Scans Identify Drug Resistance Loci in the Plasmodium Falciparum Malaria Parasite

# Sequence-based association and selection scans identify drug resistance loci in the *Plasmodium falciparum* malaria parasite

**Daniel J. Park[ab], Amanda K. Lukens[ac], Daniel E. Neafsey[a], Stephen F. Schaffner[a], Hsiao-Han Chang[b], Clarissa Valim[c], Ulf Ribacke[c], Daria Van Tyne[c], Kevin Galinsky[a], Meghan Galligan[c], Justin S. Becker[c], Daouda Ndiaye[d], Souleymane Mboup[d], Roger C. Wiegand[a], Daniel L. Hartl[abe], Pardis C. Sabeti[abe], Dyann F. Wirth[ace], and Sarah K. Volkman[acfe]**

[a]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA,[b]Harvard University, Department of Organismic and Evolutionary Biology, 52 Oxford St, Cambridge, MA 02138, USA,[c]Harvard School of Public Health, Department of Imunology and Infectious Diseases, 665 Huntington Ave, Boston, MA 02115, USA,[d]Cheikh Anta Diop University, Faculty of Medicine and Pharmacy, BP:17477, Dakar, Senegal,[f]Simmons College, School for Nursing and Health Sciences, 300 The Fenway, Boston, MA 02115, USA, and [e]These authors contributed equally

Contributed by Daniel L. Hartl

**Through rapid genetic adaptation and natural selection, the Plasmodium falciparum parasite—the deadliest of those that cause malaria—is able to develop resistance to antimalarial drugs, thwarting present efforts to control it. Genome-wide association studies (GWAS) provide a critical hypothesis-generating tool for understanding how this occurs. However, in P. falciparum, the limited amount of linkage disequilibrium (LD) hinders the power of traditional, array-based GWAS. Here, we demonstrate the feasibility and power improvements gained by utilizing whole-genome sequencing for association studies. We analyze data from 45 Senegalese parasites and identify genetic changes associated with the parasites in vitro response to twelve different antimalarials. To further increase statistical power, we adapt a common test for natural selection, XP-EHH, and utilize it to identify genomic regions associated with resistance to drugs. Using this sequence-based approach and the combination of association and selection-based tests, we detect several loci associated with drug resistance. These include the previously known signals at pfcrt, dhfr, and pfmdr1, as well as many genes not previously implicated in drug resistance roles, including genes in the ubiquitination pathway. Based on the success of the analysis presented in this study, and on the demonstrated shortcomings of array-based approaches, we argue for a complete transition to sequence-based GWAS for small, low-LD genomes like that of P. falciparum.**

GWAS | sequencing | EMMA | XP-EHH | malaria | drug resistance

Abbreviations: GWAS, genome-wide association study; LD, linkage disequilibrium; SNP, single nucleotide polymorphism; EMMA, efficient mixed-model association; XP-EHH, cross-population extended haplotype homozygosity

The malaria parasite *Plasmodium falciparum* imposes a tremendous disease burden on human societies and is responsible for 1.2 million deaths annually (1). Current efforts to eradicate malaria depend on the continued success of antimalarial drugs (2); however, the emergence of drug resistant parasites threatens to hamper global health efforts to control and eliminate the disease. Understanding the genetic basis of these adaptations will be necessary to maintain effective global health policies in the face of an ever-changing pathogen.

A key to elucidating the genetic basis of drug resistance is identifying the specific genes associated with the phenotype. In human studies of this kind, the genome-wide association study (GWAS) has overtaken the classic candidate gene approach, made affordable by the use of genotyping arrays (or SNP arrays) that measure only a subset of variants in the genome (3). This optimization is only possible because of the extensive correlation between genetic markers (called "linkage disequilibrium" or LD) in the human genome, which allows the subset of SNPs on an array to act as proxies for other markers not present; this process is known as "tagging" (4).

In *P. falciparum*, however, array-based GWAS is severely limited by the relatively short extent of LD (5–8). Lacking that correlation between genetic markers, genotyping arrays usually cannot detect associations with untyped markers, effectively limiting inferences to markers actually present on the array; even the highest density *P. falciparum* array reported to date found that LD between adjacent markers on the array was too weak for tagging in African populations (6). Consequently, current *P. falciparum* arrays cannot confidently capture all causal variants for important phenotypes.

The rapidly decreasing cost of whole-genome sequencing offers a promising solution. In principle, working with whole genome sequence allows one to directly assay all mutations segregating in the population, obviating the detection problems associated with short LD. Discovering mutations directly also avoids the ascertainment bias inherent to arrays—bias that is exacerbated when SNP discovery and genotyping are performed in different populations (9). Additionally, the small size of the *P. falciparum* genome (23Mb, roughly the size of a human exome), makes it potentially a hundred fold cheaper than whole-genome sequencing in humans. As malaria sequencing projects become cost-competitive with genotyping arrays, whole-genome sequencing has the potential to become the most effective approach to performing association studies in malaria.

Here, we test the hypothesis that whole-genome sequencing will identify SNP associations not detected by classic array-based approaches. We apply this method to identify loci in the *P. falciparum* genome that are associated with antimalarial drug resistance and compare the approach to a standard array-based GWAS. We improve the statistical power of this analysis by adapting a commonly used selection test, the cross-population extended haplotype homozygosity (XP-EHH) test (10), and utilize it as an association test for positively selected phenotypes. These approaches identify a number of candidate loci associated with anti-malarial drug resis-

**Reserved for Publication Footnotes**

tance, including genes in the ubiquitination pathway, suggesting that alteration of the parasites ability to modulate stress may contribute to evasion of drug pressure and development of resistance in *P. falciparum*.

## Results

**45 Parasite Genomes and the Absence of LD.** We chose a population in a West African region near Dakar, Senegal and culture-adapted 45 *Plasmodium falciparum* parasites recently isolated from malaria infected patients. This population is particularly relevant for these studies as it has recently been exposed to multiple, changing drug regimens as clinical resistance to traditional drugs has emerged (11). We obtained whole-genome sequence data and generated high-quality consensus base calls for an average of 83% of each genome. This produces 225,623 segregating single nucleotide polymorphisms (SNPs), of which 25,757 met our call rate and minor allele frequency criteria for further study (see Methods). Sequence-based SNP calling in *P. falciparum* is technically challenging due to its extremely AT-rich genome (12, 13). In light of this, we validated our sequence-based approach against array-based methods by using a previously described SNP array (6) to genotype 24 of the 45 isolates. Of the 74,656 SNPs assayed by the array, 4,653 meet our call rate and minor allele frequency criteria. We observe nearly perfect concordance between Affymetrix genotypes and sequence genotypes (see Methods).

Our data demonstrate that SNPs in *P. falciparum* have very little ability to tag neighboring SNPs due to the short LD in the African population from which they were sampled. While some portions of the genome exhibit significant LD, over 62% of the SNPs in the genome have no LD ($r^2 < 0.05$) between adjacent SNPs, and 87% of the SNPs have insufficient LD to tag their neighbor (Fig. 1A) using the criterion derived from human GWAS ($r^2 < 0.8$) (4). To measure tagging ability directly, we simulate genotyping arrays of various sizes by sampling random subsets of SNPs from our sequence data. We find that the simulated arrays are not able to tag a significant portion of unassayed markers, a result in stark contrast to the performance of human arrays (Fig. 1B). The tagging performance of our own Affymetrix array (tagging only 22.6% of segregating SNPs in Senegal) is even lower than simulated arrays of similar size (Fig. 1B), most likely due to population-based ascertainment biases (9) that were not modeled in our idealized approach. These findings lead us to conclude that array-based studies in *P. falciparum* will rarely be able to detect signals resulting from mutations not present on the array.

**Sequence-based Genome-wide Association Studies.** The goal of these studies is to identify genomic changes associated with changes in parasite response to antimalarial drugs, as measured in the set of 45 independent *P. falciparum* isolates. We assayed the cultured parasites for *in vitro* drug responses (measured by $IC_{50}$) to twelve standard antimalarials: amodiaquine, artemisinin, atovaquone, chloroquine, dihydroartemisinin, halofantrine, lumefantrine, mefloquine, piperaquine, primaquine, pyrimethamine, and quinine. These constitute the twelve phenotypes used in our association studies (Fig. S1). Not surprisingly, drugs with similar chemical structures (e.g. halofantrine, lumefantrine, and mefloquine) show a strong correlation in responses (Fig. S2), as has previously been observed (6, 7), and provide the opportunity for cross-validation of SNPs identified in association studies.

To test associations between SNP genotypes and drug response, we use efficient mixed-model association (EMMA). EMMA is a quantitative association approach well-suited for small sample sizes and partially inbred organisms, such as the malaria parasite (14). It is a commonly used tool among mixed-model GWAS approaches (15) and has recently demonstrated effectiveness with *P. falciparum* drug studies (6). After correcting for multiple testing (Bonferroni correction for 25,757 SNPs, $P < 2\times10^{-6}$), EMMA is able to detect a number of previously known markers of drug resistance, such as four non-synonymous SNPs in *pfcrt* (16,17) associated with chloroquine response (N75E/K, K76T, Q271E, R371I), one *pfmdr1* SNP (18, 19) associated with halofantrine, lumefantrine, and mefloquine response (N86Y), and three *dhfr* SNPs (20) associated with pyrimethamine response (N51I, C59R, S108N). We note here that, although mitochondrial and apicoplast genomes were also sequenced, no significant associations were found and the known mitochondrial mutations associated with atovaquone resistance (21, 22) were fixed in all 45 individuals for the drug-sensitive alleles (*cytochrome b* 268Y, 133M, 280G). In all, EMMA detects 34 significant SNPs associated with parasite response to five drugs (Fig. S3). Most are in or near previously known associations (8), and five are novel associations with pyrimethamine response (Supp File S6).

While these sequence-based findings validate the previously known relationship between the *pfmdr1* gene and parasite responses to halofantrine, lumefantrine, and mefloquine, it is notable that this association is not detectable by our SNP array (Fig. 2, Fig. S5), as the array lacks any markers in *pfmdr1* with a sufficiently high minor allele frequency. This exemplifies the type of association that can be missed by arrays due to limited LD. Additionally, the agreement between these three drugs at this locus provides validation of this result with respect to structurally related drugs.

**Using Haplotype-based Selection Tests for Association.** To test the hypothesis that drug resistance is largely driven by positive selection, we searched for long haplotypes associated with selection for drug resistance using the XP-EHH test (10). This selection test has not previously been used as a GWAS tool, but it is well suited for this purpose when we presume that the phenotype we are studying is under positive selection. While this assumption is not valid for most human-based GWAS for non-communicable diseases, it is very likely to be the case when studying parasite genomes for resistance adaptations to widely used drugs, which represent a strong selective pressure. Used in this way, the XP-EHH test identifies areas in the genome where resistant parasites show much longer haplotypes than sensitive parasites, indicative of recent positive selection on the resistant population. In our data, the test detects a number of signals, including *pfcrt* and *dhfr*, as well as a number of other hits spanning a total of 32 genomic regions across eleven drugs (Fig. 3, Fig. S4, Supp File S6). Seventeen of these regions are indicative of selection in the drug resistant population, whereas fifteen are consistent with selection in the drug sensitive population. With the exception of the regions containing *pfcrt* and *dhfr*, none of these loci were detected by EMMA alone.

While this approach does not detect the known *pfmdr1* locus, this is consistent with our expectations, due to the nature of the test. The N86Y mutation in *pfmdr1* confers increased susceptibility (18, 19) to many drugs when compared to the wild-type allele. As such, this SNP would not be an expected candidate for positive natural selection on a novel variant—the type of selection XP-EHH is designed to detect. Moreover, the absence of a *pfmdr1* signal from the XP-EHH test is consistent with the lack of findings in this gene from previous genomic scans for positive selection based on the REHH, iHS, and XP-EHH tests in multiple populations (5, 6, 23).

In searching for long haplotypes, the XP-EHH test typically identifies a large number of significant SNPs in close proximity to each other. These regions often span many tens of kilobases and several annotated genes. This is expected because the process of positive natural selection increases the prevalence of both the selected variant as well as of nearby variants, generating local regions of extended haplotypes. Thus, while XP-EHH strongly implicates these 32 regions as areas of phenotype-associated positive selection, by itself it is usually unable to localize the source of this selection to a specific gene. We use $P$-values from EMMA to improve signal localization by identifying the strongest signals of association within each region. This approach allows us to suggest a possible gene or mutation as a focus of phenotype-specific positive selection for each identified region (Supp File S6) and is reminiscent of earlier approaches that intersect selection and association results (23, 24).

A more comprehensive examination of the regions under drug-associated selection reveals discrete biological pathways and processes that may be particularly important as mediators of drug response in *P. falciparum* (Supp Results). The 59 genes in these 32 regions can be functionally classified as: surface molecules or transporters, genome maintenance or transcriptional regulation, metabolic enzymes including lipid metabolizers, and members of the ubiquitin proteasome system. Most surface molecule associated mutations and intergenic mutations are localized to intra-chromosomal clusters containing *var*, *rifin* and *stevor* genes; and a number of genes are found among molecules modulating ubiquitination, lipid metabolism, or folate metabolism. Members of these pathways are also represented in the large region of pyrimethamine-specific selection on chromosome 6, where it is difficult to localize the focus of selection. Collectively, these findings argue that certain biological processes in general, and genes in the ubiquitination and lipid metabolism pathways in particular, play important roles in modulating drug responses in *P. falciparum*.

## Discussion

Complete genome sequencing provides many advantages over array-based genotyping for association studies. These include the ability to directly type the causal allele, the increased detection power from increased marker density, and the ability to overcome ascertainment biases that arise when studying different populations with a fixed marker set. In *P. falciparum*, the lack of tagging ability due to the near absence of long-range LD limits the utility of arrays for association studies. Furthermore, the small genome size of *P. falciparum* brings the cost of whole genome sequencing to approximate parity with traditional genotyping arrays, and recent advances in pathogen-specific DNA-enrichment and host-specific DNA-depletion techniques for clinical samples makes the sequence-based GWAS approach more accessible and cost-effective than ever before (13, 25).

We introduce a selection-association approach based on the XP-EHH selection test. While this approach may not be appropriate for many association studies, it is sensible when the phenotype under study is under strong selection, which is likely the case for drug resistance in pathogens. As a haplotype-based test that takes advantage of multiple, adjacent SNPs, it has the advantage of being more sensitive than single-marker approaches like EMMA, given the same sample size (4). In addition to detecting new signals of drug-associated selection, we also find that the directional nature of the test statistic, a $Z$-score, provides useful information about whether the selection is associated with drug sensitivity or resistance. Consequently, we also introduce an alternative visualization of the output: a Manhattan-like plot of $Z$-scores, instead of $-\log_{10} P$-values, to illustrate the directionality of the signals (Fig. 3). In our data, we observed a tendency for many drugs (artemisnin, dihydroartemisnin, primaquine, halofantrine, lumefantrine, and mefloquine) to show highly significant signals of selection for drug sensitivity at *pfcrt*, the gene known to be responsible for chloroquine resistance (Fig. S4). While, in principle, this type of signal may result from selection towards drug sensitivity, in this particular case, it most likely results from the general pattern of anti-correlation between chloroquine and these six other drugs (Fig. S2). Additionally, the absence of a significant chloroquine sensitivity signal at *pfcrt* is consistent with reports that the return of chloroquine sensitive parasites in Africa did not result from a classic selective sweep (26). In either case, the Manhattan-like $Z$-score plots allow us to note the presence of these drug sensitivity signals while keeping them visually separate from the drug resistance signals on which we wish to focus.

Our approaches identify a significant number of loci associated with changes in drug response (Supp File S6). The strongest of these contain previously known mediators of resistance, such as the mutations in *pfcrt*, *pfmdr1*, and *dhfr*. Curation of our remaining results using a variety of gene and protein prediction algorithms and literature searches (27) point to several cellular processes and pathways of potential interest, including the ubiquitin proteasome system, lipid metabolism, and folate metabolism (Supp File S6). We argue that these findings point to biological processes used by the parasite to survive drug pressure or circumvent the action of antimalarial compounds. Other genes of interest include three ABC transporters—a class of transporters known to modulate drug responses in other organisms (28)—and genes proposed to modulate chromatin (29, 30), DNA repair (31, 32), or RNA binding (33)—pathways that have been shown to potentially be altered in response to drug pressure.

A number of the signals of recent positive selection are unique to pyrimethamine-resistant parasites. While the known resistance locus, *dhfr*, is present among these, there are even stronger signals of pyrimethamine-associated selection on chromosome 6 and chromosome 12. The region on chromosome 6 contains two previously uncharacterized genes proposed to participate in folate metabolism (PFF1360w and PFF1490w), as well as six genes acting as either chaperones or in ubiquitination (PFF1365c, PFF1485w, PFF1445c; PFF1415c; and PFF1505w), and three molecules likely to modulate lipid metabolism (PFF1350c, PFF1375c-a/b, and PFF1420w). In the chromosome 12 region, the XP-EHH test produces significant $P$-values for eight SNPs over a 15kb region spanning five adjacent genes. The extended haplotypes surrounding these SNPs continue even further, spanning 28kb and fourteen genes in total (Fig. 4A). These results present challenges for experimental validation, as the goal of association studies is to generate a small number of testable hypotheses about molecular mechanisms. Fortunately, the use of EMMA $P$-values in this region can assist in localizing the signal. We find that the strongest EMMA SNP coincides with the strongest XP-EHH SNP, which is a non-synonymous mutation in PFL2100w, a putative ubiquitin conjugating enzyme (E2) (Fig. 4B). Additionally, a significant, pyrimethamine-specific selection signal on chromosome 8 is entirely contained within MAL8P1.23 (a putative HECT ubiquitin ligase E3) (Supp File S6), another gene in the ubiquitin-mediated pathway (34). Given the role of this pathway in directing protein degradation and recycling, it is possible that alterations in these genes create changes in stress responses or protein

turnover of key resistance modulators that allow the parasite to survive under drug pressure.

The evolution of drug resistance in the natural setting is likely to be a multistep process and our work potentially identifies key pathways involved in this process. Field-based evidence has demonstrated a reduced fitness for drug resistant parasites in the absence of drug pressure and laboratory-based work has demonstrated the relative fitness of different mutational changes in target enzymes. Our findings point to potential compensatory mutations in a pathway related to protein stability and turnover and it is tempting to speculate that such adaptations enable the "expression" of a resistant phenotype, such as has been observed in yeast (35). Although molecular approaches are required to validate the role of this pathway in modulating drug response, these results demonstrate the potential for sequence-based GWAS approaches to identify pathways, in addition to individual genes, that may be responsible for the phenotype of interest.

Ultimately, all association results require experimental validation and follow-up work to explore possible mechanisms of action. Association studies, even in their ideal form, simply generate hypotheses based on correlations. However, improved methods for association studies can significantly reduce the necessary validation work by reducing false positive rates, increasing study detection power, and improving localization ability. This study successfully pilots the use of whole-genome sequence data for association studies in malaria and it demonstrates significant advantages in detection power over array-based studies. We strongly recommend that future association studies in low-LD, small-genome organisms adopt the sequence-based GWAS approach as well, given the relative costs. We additionally demonstrate the effectiveness of the XP-EHH selection test as an association test for phenotypes under positive selection. Finally, we combine data from both tests to localize long signals and reduce the number of hypotheses for follow-up validation. This combined approach identifies more candidate loci than with single-marker tests alone.

## Materials and Methods

**Sequencing.** Parasites were obtained from patients with uncomplicated mild malaria in Senegal from 2001 to 2009 under ethical approval with informed consent for the study. Parasites were culture adapted by standard methods (36) and genomic DNA was extracted from 45 single-clone samples. Samples were determined to be monoclonal and genetically distinct by a 24 SNP molecular barcode (37). Genomic DNA was sequenced using Illumina Hi-Seq machines. The first 12 parasites were sequenced with 76bp single-end reads and the remaining 33 were sequenced with paired-end reads ranging from 76bp to 101bp in length. The median sequence coverage depth was 144.8X after alignment (ranging from 32X to 400X). Reads were aligned with BWA v0.5.9-r16 against the 3D7 reference assembly (PlasmoDB v7.1). A consensus sequence was called for each strain using the GATK Unified Genotyper v1.2.3-g61b89e2 (38) with the following parameters: `-A AlleleBalance -stand_emit_conf 0 --output_mode EMIT_ALL_SITES`. Bases were then removed if they exhibited poor quality (GQ less than 30 or QUAL less than 60) or if they called a heterozygous genotype. This left consensus calls for 56-91% of the genome (83% median) for each of 45 individuals. Of these sites, 225,623 positions are polymorphic among the 45 individuals. Of these SNPs, only 25,757 had genotypes in at least 36 individuals (80% call rate) and were non-singletons (i.e. minor allele count $> 1$ or minor allele frequency $> 4\%$). All analyses are based on this set of 25,757 SNPs. SNP data is available in dbSNP as batch Pf_0004 from submitter BROAD-GENOMEBIO. SNPs are being processed at PlasmoDB (27) for release later this year. SNP data can also be found in Supp File S9. Consensus calls for the whole genome are available in Supp File S10.

Principal component analysis was conducted using the program *SMARTPCA* (39) in the EIGENSOFT 3.0 package. We applied a local LD correction (nsnpldregress = 2) and found no significant eigenvectors in the population.

**Tagging Analysis.** Tagging analysis in Figure 1B was generated by using PLINK (40) to find tagging SNPs for each SNP that were within 10kb and at least $r^2 >= 0.8$. We then simulated genotyping arrays by randomly sampling subsets of SNPs of varying subset sizes and calculating the fraction of total SNPs that are tagged by the subset. We first reduced the sequence data to 40 random individuals to simulate ascertainment bias against low allele-frequency markers, then randomly sampled markers that were still polymorphic among the smaller population size to simulate a genotyping array. We simulated 19 different array sizes, ranging from 5% of the sequenced SNPs (1,227) to 95% of the sequenced SNPs (22,087). 200 simulations per array size were run and the result was highly consistent: 95% confidence intervals were too small to visualize on the figure. Simulations for the human genome were based on 60 diploid individuals of European descent (CEU) from Hapmap release 23a. Each iteration chose 54 random individuals to simulate ascertainment bias, filtered SNPs to an 80% call rate and to non-singletons. Our Affymetrix array was able to tag 5,508 SNPs in our sequence data using the 4,894 SNPs on the array that overlapped with the 25,757 SNPs in our sequence data (open triangle in Fig 1b). Histograms in Figure 1A are binned into 20 evenly spaced bins of $r^2$ from 0 to 1. The plot is normalized such that the sum of all bars in each histogram is equal to 1 to show the relative proportions of SNPs in each bin. Simulation data is provided in Supp File S7.

**Drug Assays.** Drug assays were performed as described (41) with slight modifications for 384-well format (Supp Methods). The range of drug concentrations are shown in Figure S1, and the $\mathrm{IC}_{50}$ data is provided in Supp File S8. Raw input data for all association tests is provided in Supp File S9.

**EMMA.** Single marker association tests were run using EMMA (14). Since not all drugs have complete phenotype data for all 45 individuals, SNPs are additionally filtered to those that met our previous call rate and minor allele criteria among the subset of samples for which drug data exists. This results in 23,000 to 25,180 SNPs for any given drug. $\log_{10}(\mathrm{IC}_{50})$ values were used for this quantitative test. Biological replicates of drug data were presented to EMMA as multiple individuals from the same genetic strain. This allows EMMA to use the additional data to discern heritable phenotypic variance from non-heritable variance (15), and mimics the use of clonally identical parasites in other studies (42, 43). Significance was defined as SNPs that exceeded a Bonferroni-corrected threshold of $P < 0.05$ while also surviving 60% of jackknife simulations. EMMA results were jackknifed by performing 200 random subsets of 38 samples and requiring an FDR-corrected significance of $Q < 0.1$. SNPs that passed this threshold in 60% of jackknife simulations were considered to be robust against false positives due to small sample size effects.

**XP-EHH.** Selection-association tests were run using the cross population extended haplotype homozygosity test (XP-EHH) (10). Each drug defined a partitioning of samples into two "subpopulations" ("sensitive" and "resistant") based on cutoffs shown in Figure S1 and Supp File S8 (Supp Methods). XP-EHH requires a recombination map as input, which we constructed with LDhat v2.1 (44) (Supp Methods). XP-EHH also requires fully imputed genotypes. Imputation was performed using PHASE 2.1.1 (45), producing 29,605 non-singleton SNPs (Supp Methods).

XP-EHH computes a significance value for each SNP in the genome, assuming that SNP comprises the haplotype "core" of selection. Because the test identifies long haplotypes, it results in a large number of genome-wide significant SNPs (defined by Bonferroni-corrected $P < 0.05$) in clustered stretches of the genome. We reduced the set of significant SNPs to a set of significant genomic regions by taking each significant core SNP, computing a window around each one where EHH decayed to 0.05, and merging overlapping windows. This resulted in a smaller list of significant regions for each drug (Supp File S6). Regions were further filtered by removing those which did not contain at least one core SNP that survived 50% of jackknife simulations. XP-EHH results were jackknifed by performing 200 random subsets of 38 samples and requiring a Bonferroni-corrected significance of $P < 0.1$.

**Genotyping Arrays.** A subset of 25 parasites was also hybridized to an Affymetrix array containing 74,656 markers (6). SNPs were called using BRLMM-P from Affy Power Tools v1.10.2 and filtered according to the same methods as Van Tyne, et al (6), resulting in 15,075 validated SNPs, 8,778 of which were polymorphic among the 25 individuals from Senegal. SNP coordinates were converted from PlasmoDB v5.0 coordinates to v7.1 coordinates using whole genome nucmer alignments (46). Concordance between array and sequencing data was measured for the set of markers in which genotype calls existed by both methods. For 24 samples, nearly perfect concordance between Affymetrix genotypes and sequence genotypes was observed for the 24 samples (averaging 99.2% concordance, with all 24 samples above 98.2% concordance). This level of concordance is similar to what is observed with technical replicate hybridizations of the same DNA sample (6). One sample, SenP19.04.c, reported a 28.2% mismatch rate, suggestive of a sample identification error, and was removed from the analysis. EMMA analyses were run on the array data using the same filters and procedures as for sequence data described above, utilizing 4,514-4,653 SNPs per drug phenotype. Results are shown in Fig S3. Array data for these 24 samples are in Supp File S9.

**ACKNOWLEDGMENTS.** We thank the sample collection team in Senegal, including Younouss Diedhiou, Lamine Ndiaye, Amadou Moctar Mbaye, Baba Dieye, Moussa Dieng Sarr, Papa Diogoye Sene, and Ngayo Sy. We thank the technical staff at HSPH who maintained parasite cultures, including Kayla Barnes, Dave Rosen, Kate Fernandez, and Gilberto Ramirez. We thank members of the Sabeti lab for a careful review of our manuscript, including Kristian Andersen, Chris Edwards, Chris Matranga, Rachel Sealfon, Jesse Shapiro, Ilya Shlyakhter, Matt Stremlau, and Shervin Tabrizi.

We acknowledge contributions made to the community database, PlasmoDB.org, that facilitated biological curation of candidate genes presented in this work.

This study is supported by the Bill and Melinda Gates Foundation, National Institutes of Health (Grant: 1R01AI075080-01A1), Ellison Medical Foundation, Exxon-Mobil Foundation, NIH Fogarty, NIAID, and Broad SPARC. DJP is supported by an NSF Graduate Research Fellowship. PCS is supported by fellowships from the Burroughs Wellcome and Packard Foundations.

1. Murray CJL, *et al.* (2012) Global malaria mortality between 1980 and 2010: a systematic analysis. *The Lancet* 379:413–31.
2. malERA Consultative Group on Drugs (2011) A research agenda for malaria eradication: drugs. *PLoS Medicine* 8:e1000402.
3. Altshuler DM, Daly MJ, Lander ES (2008) Genetic mapping in human disease. *Science* 322:881–8.
4. de Bakker PIW, *et al.* (2005) Efficiency and power in genetic association studies. *Nature Genetics* 37:1217–23.
5. Mu J, *et al.* (2010) Plasmodium falciparum genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nature Genetics* 42:268–271.
6. Van Tyne D, *et al.* (2011) Identification and functional validation of the novel antimalarial resistance locus *pf10_0355* in *plasmodium falciparum*. *PLoS Genetics* 7:e1001383.
7. Yuan J, *et al.* (2011) Chemical genomic profiling for antimalarial therapies, response signatures, and molecular targets. *Science* 333:724–9.
8. Volkman SK, Neafsey DE, Schaffner SF, Park DJ, Wirth DF (2012) Harnessing genomics and genome biology to understand malaria biology. *Nature Reviews Genetics* 13:315–328.
9. Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment biases in snp chips affect measures of population divergence. *Molecular Biology and Evolution* 27:2534–47.
10. Sabeti PC, *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913–8.
11. Mouzin E, Thior PM, Diouf MB, Sambou B (2010) Focus on senegal. Tech. Rep. 4, World Health Organization.
12. Oyola SO, *et al.* (2012) Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. *BMC Genomics* 13:1.
13. Melnikov A, *et al.* (2011) Hybrid selection for sequencing pathogen genomes from clinical samples. *Genome Biology* 12:R73.
14. Kang HM, *et al.* (2008) Efficient control of population structure in model organism association mapping. *Genetics* 178:1709–23.
15. Price AL, Zaitlen NA, Reich DE, Patterson N (2010) New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* 11:459–63.
16. Fidock DA, *et al.* (2000) Mutations in the p. falciparum digestive vacuole transmembrane protein pfcrt and evidence for their role in chloroquine resistance. *Molecular Cell* 6:861–71.
17. Wootton JC, *et al.* (2002) Genetic diversity and chloroquine selective sweeps in plasmodium falciparum. *Nature* 418:320–3.
18. Duraisingh MT, *et al.* (2000) The tyrosine-86 allele of the pfmdr1 gene of plasmodium falciparum is associated with increased sensitivity to the anti-malarials mefloquine and artemisinin. *Molecular and Biochemical Parasitology* 108:13–23.
19. Nkhoma S, *et al.* (2009) Parasites bearing a single copy of the multi-drug resistance gene (pfmdr-1) with wild-type snps predominate amongst plasmodium falciparum isolates from malawi. *Acta tropica* 111:78–81.
20. Nair S, *et al.* (2003) A selective sweep driven by pyrimethamine treatment in southeast asian malaria parasites. *Molecular Biology and Evolution* 20:1526–36.
21. Kessl JJ, Meshnick SR, Trumpower BL (2007) Modeling the molecular basis of atovaquone resistance in parasites and pathogenic fungi. *Trends in Parasitology* 23:494–501.
22. Dong CK, *et al.* (2011) Identification and validation of tetracyclic benzothiazepines as plasmodium falciparum cytochrome bc1 inhibitors. *Chemistry & Biology* 18:1602–1610.
23. Cheeseman IH, *et al.* (2012) A major genome region underlying artemisinin resistance in malaria. *Science* 336:79–82.
24. Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET, Pritchard JK (2009) Gene expression levels are a target of recent natural selection in the human genome. *Molecular Biology and Evolution* 26:649–658.
25. Venkatesan M, *et al.* (2012) Using cf11 cellulose columns to inexpensively and effectively remove human dna from plasmodium falciparum-infected whole blood samples. *Malaria Journal* 11:41.
26. Laufer MK, *et al.* (2010) Return of chloroquine-susceptible falciparum malaria in malawi was a reexpansion of diverse susceptible parasites. *Journal of Infectious Diseases* 202:801–8.
27. Aurrecoechea C, *et al.* (2009) Plasmodb: a functional genomic database for malaria parasites. *Nucleic Acids Res* 37:D539–43.
28. Leprohon P, Légaré D, Ouellette M (2011) Abc transporters involved in drug resistance in human parasites. *Essays Biochem* 50:121–44.
29. Cui L, Miao J (2010) Chromatin-mediated epigenetic regulation in the malaria parasite plasmodium falciparum. *Eukaryotic Cell* 9:1138–49.
30. Coleman BI, Duraisingh MT (2008) Transcriptional control and gene silencing in plasmodium falciparum. *Cell Microbio* 10:1935–46.
31. Castellini MA, *et al.* (2011) Malaria drug resistance is associated with defective dna mismatch repair. *Mol Biochem Parasitol* 177:143–7.
32. Tarique M, Satsangi AT, Ahmad M, Singh S, Tuteja R (2012) Plasmodium falciparum mlh is schizont stage specific endonuclease. *Mol Biochem Parasitol* 181:153–61.
33. Meng X, *et al.* (2012) Cytoplasmic metadherin (mtdh) provides survival advantage under conditions of stress by acting as rna-binding protein. *J Biol Chem* 287:4485–91.
34. Ponts N, *et al.* (2008) Deciphering the ubiquitin-mediated pathway in apicomplexan parasites: a potential strategy to interfere with parasite virulence. *PLoS ONE* 3:e2386.
35. Jarosz DF, Lindquist S (2010) Hsp90 and environmental stress transform the adaptive value of natural genetic variation. *Science* 330:1820–4.
36. Trager W, Jensen JB (1976) Human malaria parasites in continuous culture. *Science* 193:673–5.
37. Daniels R, *et al.* (2008) A general snp-based molecular barcode for plasmodium falciparum identification and tracking. *Malaria Journal* 7:223.
38. McKenna A, *et al.* (2010) The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research* 20:1297–303.
39. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics* 2:e190.
40. Purcell S, *et al.* (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81:559–75.
41. Plouffe D, *et al.* (2008) In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen. *Proc Natl Acad Sci USA* 105:9059–64.
42. Anderson TJC, *et al.* (2010) Inferred relatedness and heritability in malaria parasites. *Proc R Soc B* 277:2531–40.
43. Anderson TJC, *et al.* (2010) High heritability of malaria parasite clearance rate indicates a genetic basis for artemisinin resistance in western cambodia. *Journal of Infectious Diseases* 201:1326–30.
44. McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160:1231–41.
45. Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics* 73:1162–9.
46. Kurtz S, *et al.* (2004) Versatile and open software for comparing large genomes. *Genome Biology* 5:R12.

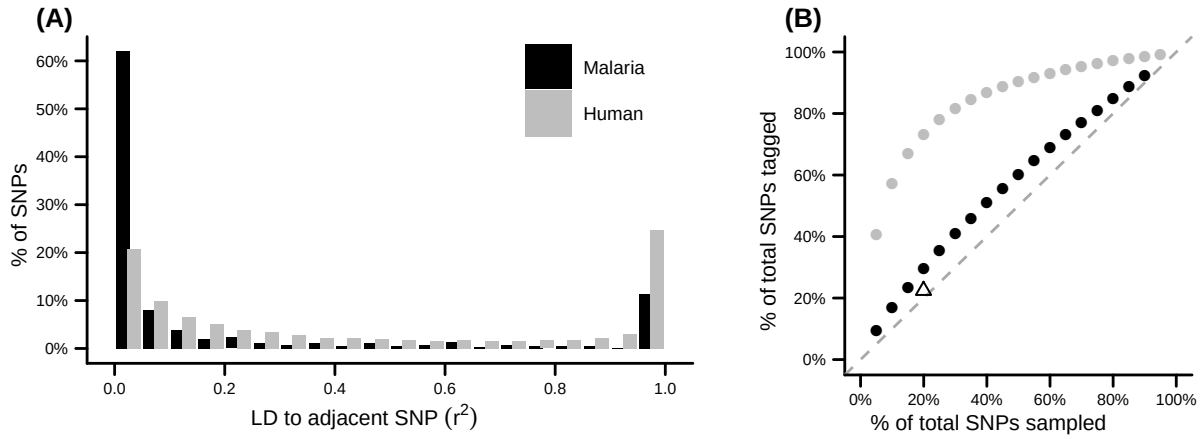


**Fig. 1.** Simulated *P. falciparum* arrays are unable to tag SNPs not present on the array. **(A)** A histogram of LD between adjacent SNPs from sequenced *P. falciparum* (black). The vast majority of markers have little to no LD with their neighbors (62% of SNPs have $r^2 \leq 0.05$, 76% have $r^2 \leq 0.2$, and 87% have $r^2 \leq 0.8$). This contrasts with human studies where much more of the genome shows moderate to strong LD between neighboring SNPs (gray). **(B)** Simulated genotyping marker sets of various sizes are plotted against the percentage of the entire sequenced marker set that they are able to tag (with $r^2 \geq 0.8$). The dashed, identity line depicts the theoretical scenario where all SNPs are in complete linkage equilibrium and no SNP tags another. Since this is true of 87% of SNPs in the malaria sequence data, the increase is almost linear (black dots). This contrasts with the array tagging performance seen in human studies (gray dots), where only a small fraction of markers are needed to tag the bulk of the genome—a principle that array-based GWAS depends on. The open triangle depicts the actual performance of the Affymetrix-based Broad Institute *P. falciparum* SNP array (6).

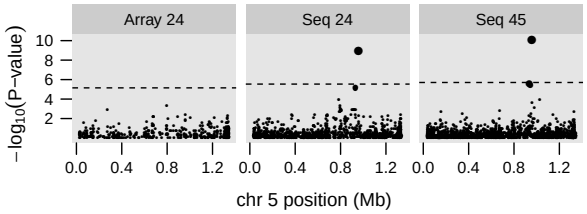


**Fig. 2.** Mefloquine association signals around the known drug resistance locus *pfmdr1*. EMMA results are shown for all of chromosome 5 with $P$-values for each SNP on a $-\log_{10}$ scale against physical position. The array-based study (Array 24) does not detect any association at the known *pfmdr1* locus due to a lack of marker coverage within the gene and sufficient LD around the gene. The sequence-based study with the same 24 samples (Seq 24) detects the expected hit at 0.96Mb. Including all samples from the sequence-based study (Seq 45) increases the strength of this signal. The dashed line indicates the Bonferroni-corrected significance threshold ($P = 0.05$, genome-wide SNP counts are 7,068, 17,278, and 25,159 respectively).

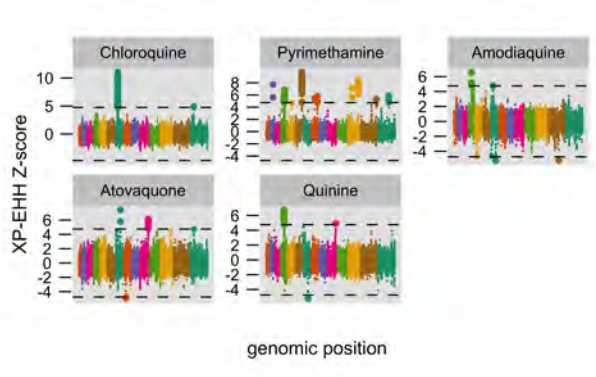


**Fig. 3.** Significant signals of drug-associated selection across five antimalarial drugs. XP-EHH results are shown using a Manhattan-inspired plot, with SNP $Z$-scores plotted against genomic position, with each chromosome colored separately. Positive $Z$-scores suggest selection in drug resistant parasites, negative $Z$-scores suggest selection in sensitive parasites. The dashed lines indicate the two-sided Bonferroni significance thresholds ($P = 0.025$ and $0.975$). Only drugs with significant hits are shown here, $Z$-score and quantile-quantile plots for all drugs are shown in Fig. S4.

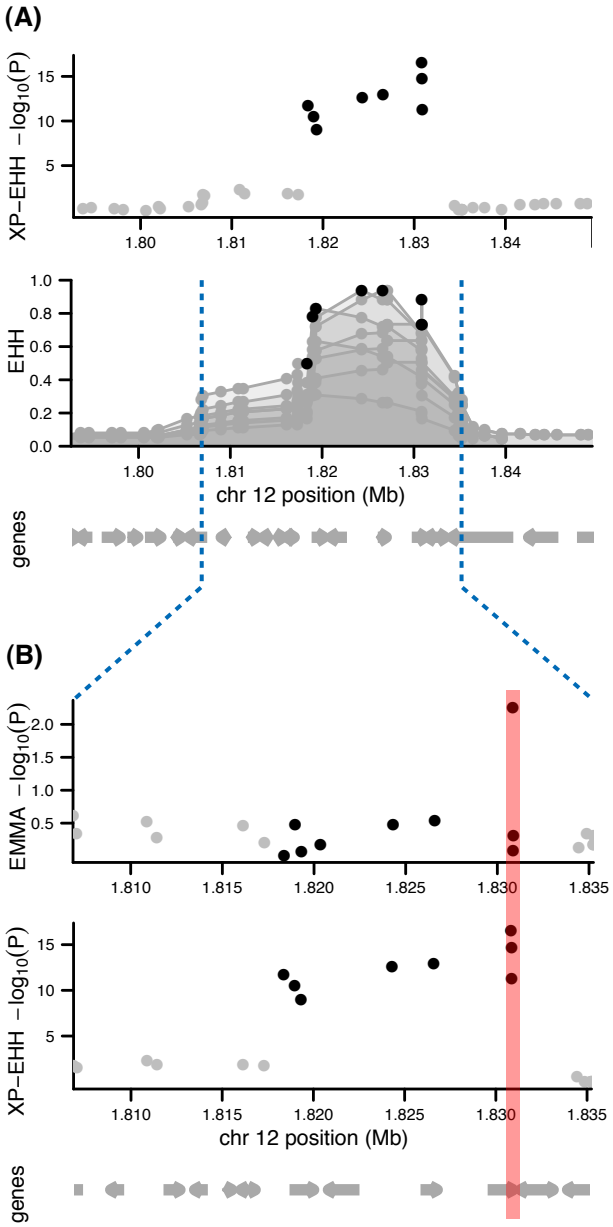


**Fig. 4.** Localizing the pyrimethamine-associated selection signal on chromosome 12. **(A) Defining the region:** XP-EHH identifies eight genome-wide significant SNPs in close proximity on chromosome 12. Each of these eight SNPs represents the center of an area of extended haplotype homozygosity, as measured by the EHH statistic. Haplotype decay for resistant parasites is plotted for each of these eight SNPs, which defines a larger region from 1.807Mb to 1.835Mb in which the causal mutation may exist. This region spans 28kb and 14 genes. **(B) Localizing the signal:** focusing within this region, we utilize single-marker association signals from EMMA to localize the signal. The most significant EMMA SNP coincides with the most significant XP-EHH SNP and localizes to an E398D mutation in PFL2100w (ubiquitin conjugating enzyme E2).